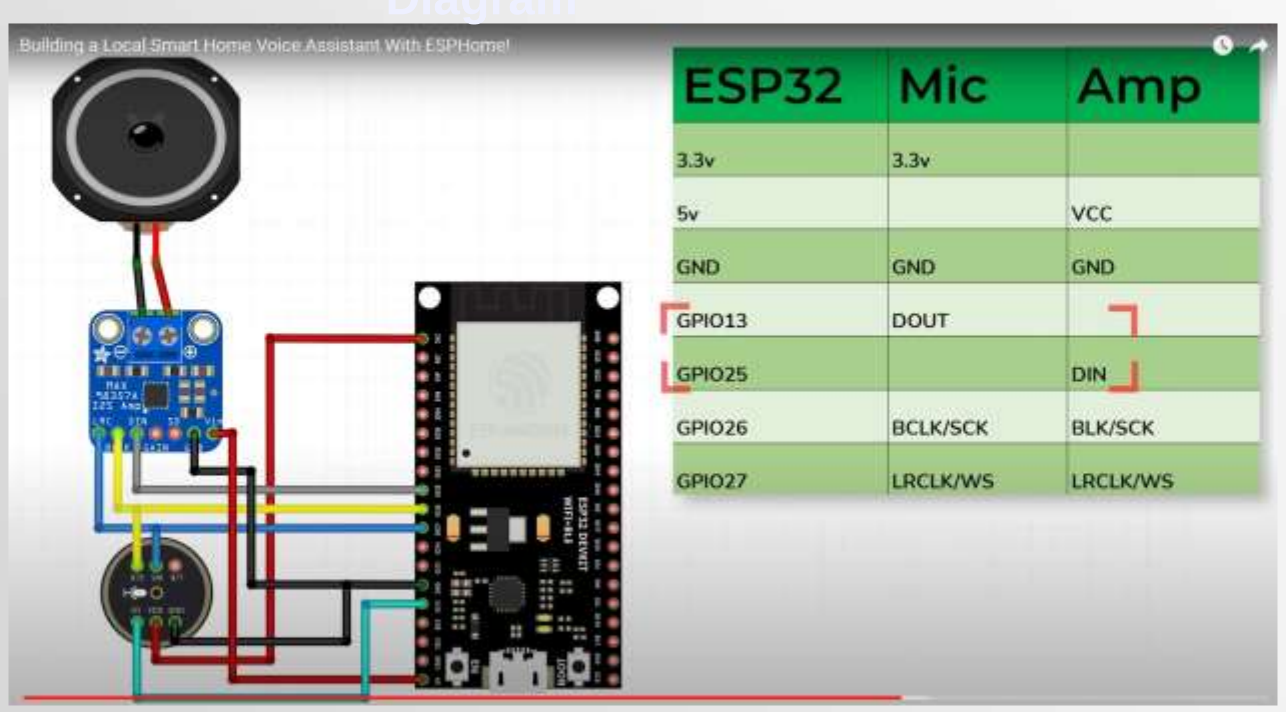
Real-Time Multilingual Voice Translator with ESP32 and AI Integration
AIOT & it's applications

BY: V. ABHIVARUN , CHARUL PAREEK , DIVYANSHU SHEKHAR, KRUTHIKA PRIYA CHANDAN

Overview

This is an offline AI-based speech-to-speech translator project centering on the ESP32 microcontroller that can facilitate real-time communication in multiple languages without being dependent on the internet. It records spoken data via an I2S microphone, performs on-device audio processing with Vosk to perform speech-to-text, translates the transcribed text using Argos Translate, and generates output speech using pyttsx3, all on a local server written in Flask. The translated audio output is heard via a speaker, which is connected to the ESP32, which is great since it requires no connectivity or can be used in low connectivity areas such as rural zones, traveling, or schools. It can be highlighted that its full offline support, cheap hardware, modular design, and multilingual support (e.g., English to Hindi/Telugu) are its important qualities. future features might include LLMs to respond in context or a mobile app to pick the language.

Schematic Diagram



Components And Modules

Software Modules

File Name	Function
record_audio_final.py	Captures user's voice using **I2S mic**, saves it as Recording.wav on ESP32
speech_to_text.py	Sends recorded audio to **Vosk STT Flask server**, retrieves **transcribed text**
text_to_speech.py	Sends transcription to **Translate+TTS Flask server**, receives **audio response**
play_audio_final.py	Plays the translated voice via **MAX98357A DAC and speaker**

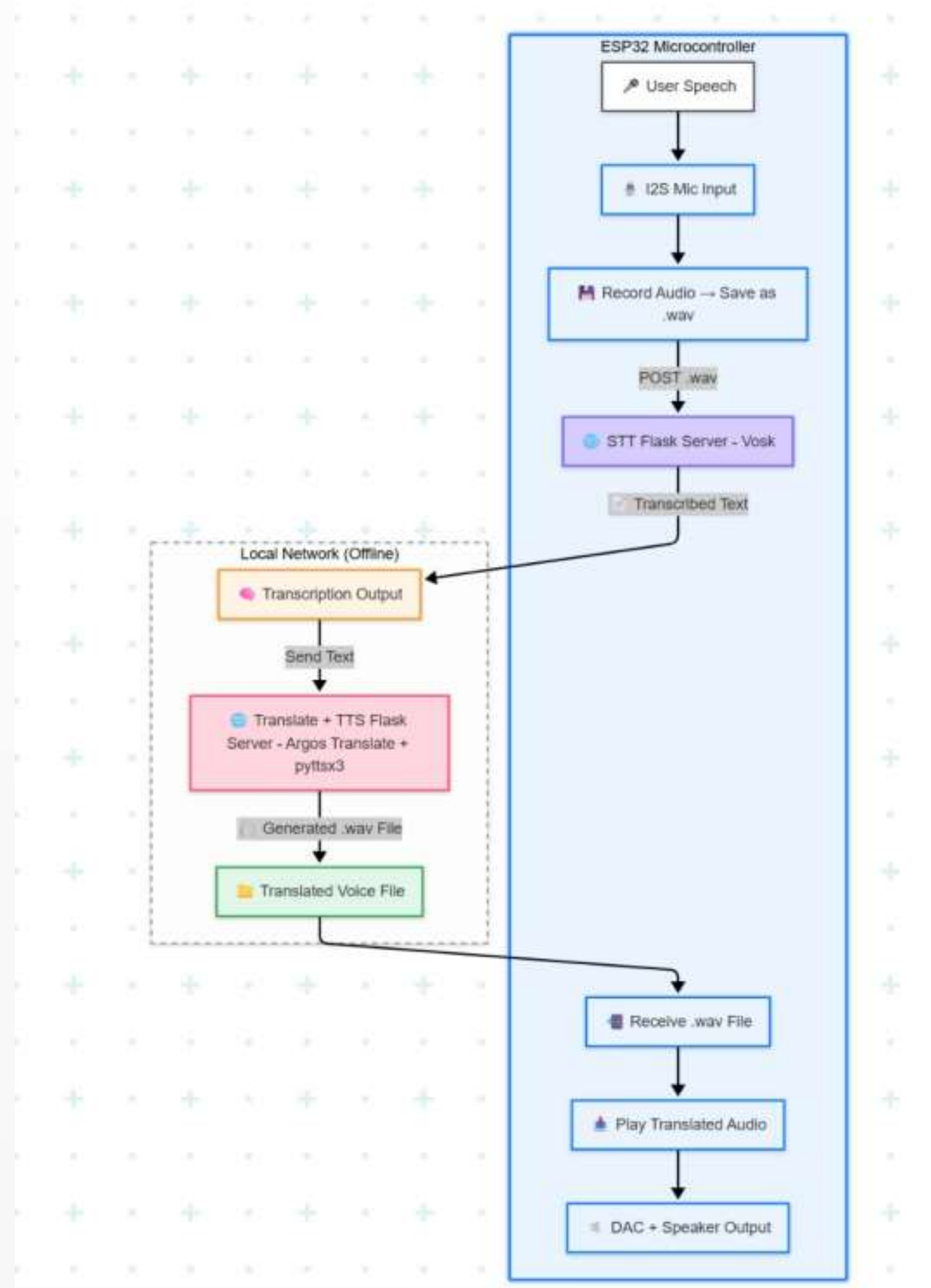
Backend Flask Servers

Server	Function
translate_speak_server.py	Uses **Argos Translate + pyttsx3** for offline TTS
stt_server.py / server.py	Runs **Vosk STT** to convert WAV to text

Hardware Components

Component	Function
ESP32 Microcontroller	Main controller; handles Wi-Fi, I2S audio input/output, and server communication
I2S Microphone (INMP441)	Captures high-quality audio from user's voice using I2S interface
MAX98357A DAC Module	Converts digital audio to analog signals for speaker playback
Speaker	Outputs the translated speech audibly to the user

FlowChart



Future Enhancements

1. All-in-One Embedded Solution (Without Laptop/RPi)
Integrate models directly on a more powerful embedded board (e.g., Raspberry Pi Zero 2 W or Jetson Nano) to eliminate the need for external PC.

2. Wake Word Detection or Language Switching Interface
Implement local wake word detection (e.g., "Hey Translator") for more intuitive, hands-free interaction.

3. Noise Reduction Techniques
Add signal filtering or lightweight ML-based noise suppression for better mic performance in the field.

4. Language Expansion
Add support for more regional and international languages (e.g., Tamil, Marathi, Bengali, Spanish) by downloading/installing new Argos Translate models.

RESULT

Detecting text using esp32

Pipeline Execution Results

Stage	Time (ms)	Time (s)
Audio Recording	3010	3,01
Speech-to-Text (STT) Using Vosk STT Server	34139	34,14
Translation + TTS Argos Translate + pyttsx3	1930	1,93
Audio Playback Through I2S DAC+Speaker	2339	2,34
Total Time	41420 s	41,42 s

Speech Translation Result

Captured Input	Translated
Hello...	ఎలో...

- Fille name: Recording.wav
- Output: synthesized_audio.wav
- Mic: I2S (ESP32)
- Output: MAX98357A DAC + Speaker

INPUT

```
>>> %Run -c $EDITOR_CONTENT

MPY: soft reboot
   Recording audio...
Recording for 3.0 seconds
Finished Recording
 Audio recording completed.
 Recording time: 3010 ms

 Sending audio to Vosk STT server...
{'message': 'Chunk received'}
{'message': 'Chunk received'}
{'message': 'Chunk received'}
{'message': 'Chunk received'}
{'message': 'Chunk received'}
{'message': 'Chunk received'}
{'message': 'Chunk received'}
{'message': 'Chunk received'}
{'transcript': 'hello how are you'}
```

OUTPUT

```
Transcription result:
hello how are you

STT time: 34139 ms

 Translating and fetching TTS...
 Playing translated speech...
Starting
Done
 Audio playback time: 2339 ms

 TTS + download time: 1930 ms
 Total pipeline time: 41420 ms
>>>
```

CONCLUSION

This project demonstrates a **low-cost, offline-capable** speech translator using ESP32 and edge AI, bridging embedded systems with real-time language processing. Future work focuses on **optimizing latency, expanding languages, and reducing server dependency**—paving the way for **standalone, battery-powered translators**. Ideal for education, travel, and IoT, it highlights the potential of **on-device AI for accessible, privacy-first communication**

References

- **P. Kshirsgar, "Advances in Offline Speech Processing," Springer Nature, Singapore, 2020. This is highly relevant as it directly covers offline speech processing, which matches your project's core innovation of working without internet connectivity.**
- **A. Ismail, "Deep Learning-Based Speech Translation and Sustainability," Sustainability, vol. 12, p. 2403, 2020. This reference is valuable because it focuses on speech translation and edge AI sustainability, which relates to your low-power ESP32 implementation.**
- **G. K. K. Sanjivani S. Bhabad, "An Overview of Technical Progress in Speech Recognition," International Journal of Advanced Research in Computer Science and Software Engineering, 2013. This provides fundamental insights into speech recognition technology, supporting the Vosk/STT component of your project.**
- **Kaveri Kamble, Ramesh Kagalkar, "A Review: Translation of Text to Speech Conversation for Hindi Language," International Journal of Science and Research (IJSR), 2012. This is particularly relevant as it discusses text-to-speech conversion for Hindi, one of your target languages.**
- **Chethan, "Offline Voice-Based Applications," International Journal of Engineering Research and Technology (IJERT), vol. 2, no. 5, 2017. This reference is important as it explores offline voice systems, which is crucial for your project's internet-independent operation.**

Acknowledgments

We thank our mentors for their guidance and the open-source community for valuable resources. Special thanks to our friends and family for their support and encouragement.